統合と推論：IIT、FEP、そして深層学習の未来に関する統一的分析

第I部 心と機械に関する基礎理論

第1章 統合情報理論：内的因果力の計算体系

統合情報理論（Integrated Information Theory, IIT）は、その現象学的な出発点から物議を醸す存在論的な結論に至るまで、意識に関する最も包括的かつ厳密な理論の一つとして位置づけられています。本章では、後続の議論の基盤として、IITの完全な理論的・数学的基礎を確立します。

1.1 現象学の公理：意識体験の本質的特性の定義

IITは、考えうるあらゆる意識体験が持つ、自明かつ反論の余地のない特性を「公理」として定義することから始まります[1]。これは、理論が物理的なメカニズムの説明に進む前の、現象そのものの記述です。5つの中心的な公理は以下の通りです。

- 内在性（**Intrinsicality**）：意識体験は、それ自体のために、その内在的な視点から存在する[1]。
- 構成性（**Composition**）：意識体験は構造化されており、区別（distinctions）とそれらの関係（relations）から構成される[1]。
- 情報（**Information**）：意識体験は特異的である。それは「そのようである」という特定のものであり、それによって膨大な数の他の可能な体験から区別される[1]。
- 統合（**Integration**）：意識体験は単一であり、それ以上還元できない。例えば、赤い正方形の体験において、色と形を分離することはできない[1]。
- 排他（**Exclusion**）：意識体験は、その内容と時空間的な粒度において限定的である。そ

れは、そこに含まれるものを含み、それ以上でもそれ以下でもない [1]。

これらの公理は、IITが次にメカニズム的に説明しようとする現象学の完全かつ交渉の余地のない記述を形成することを意図しています [1]。

1.2 存在の要請：主観性から物理システムへ

IITの核心は、現象学からメカニズムへの「飛躍」にあります。これは、前述の公理を、物理的な基盤が満たすべき必要十分条件に関する「要請」へと変換するプロセスです [1]。対応する要請は以下の通りです。

- 存在（**Existence**）**:** システムは因果力、すなわち「違いを受け、違いを生む」能力を持たなければならない [4]。これはプラトンのエレア派の存在原理と密接に関連しています [4]。
- 構成、情報、統合、排他**:** これらの要請は公理を反映しており、物理システムが相互作用する要素から構成され、それらが還元不可能な因果構造（概念構造）を特定し、その構造が限定的かつ最大であることを要求します [4]。

この文脈で、システムはその遷移確率行列（Transition Probability Matrix, TPM）によって完全に特徴づけられ、これがシステムの因果力を記述します [1]。

1.3 Φ指標：還元不可能な統合の定量的尺度

IITは、統合情報量、通称「ファイ」（Φ）と呼ばれる指標を提案します [1]。

Φは、システム全体として特定される因果構造が、その部分の総和にどれだけ還元できないかを定量化するものです [8]。

Φの計算プロセスは、システムを考えうるすべての最小情報分割（Minimum Information Partition, MIP）で分割し、接続を「仮想的に切断」し、失われる情報量を測定することを含みます [10]。最も情報損失が少ない分割が、そのシステムの

Φ値を定義します。

この計算から、「コンプレックス」（complex）という概念が導入されます。これは、Φが局所的に最大となる要素のサブセットであり、単一の意識体験の基盤となります [1]。さらに、「最大存在

の原理」によれば、重なり合う候補システムの中で、最も還元不可能な因果力（

Φmax）を持つものだけが、意識を持つ実体として真に存在するとされます [4]。

1.4 Φ構造と体験の質

IITは、意識の「量」（Φ）を定量化するだけでなく、その「質」（what it is like）を特徴づけることも目指しています。これは、「Φ構造」または「概念構造」と呼ばれるものを通じて達成されます。これは、コンプレックスによって特定される因果レパートリー（概念）の完全な集合を、高次元空間に展開したものです [1]。

IITの「説明的同一性」とは、現象的な体験がこのΦ構造と「同一」であるという主張です。体験がどのように感じられるかは、それに対応する概念構造の幾何学的形状と同一であるとされます [1]。

1.5 批判的分析：汎心論、反証不可能性、そして「存在の大きな分断」

IITはその理論的射程の広さから、重大な論争を引き起こしてきました。特に2023年には、反証不可能な疑似科学であるとの批判もなされています [1]。

理論の含意の一つは汎心論的、あるいは準汎心論的な帰結です。多くのシステムでΦがゼロでない値を取りうるため、IITは意識が程度の問題であり、宇宙に広く存在する可能性を示唆します [5]。これは、例えばロジックゲートのネットワークが意識を持つといった、直感に反する結論を導きます [15]。

さらに、IIT 4.0で提唱された「存在の大きな分断」という存在論的主張は、最大のΦを持つ意識的な実体（コンプレックス）のみが「真の」あるいは「内在的な」存在を持ち、他のすべての物体は観察者の視点から相対的にのみ存在すると主張します [16]。これは観念論や二元論の一形態として批判されています [17]。

より洗練された批判として、IITを2つのレベルに分ける見方があります [15]。レベル1は、システムの複雑性を測る有用な定量的理論としての側面です。レベル2は、この複雑性を現象的意識と同一視する「疑わしい飛躍」です。この見方では、レベル1はシステムの「自律性」を定義するものの、それは意識を持つための基準としてはるかに低いものだと主張されます。

IITの計算論的観点からの主要な貢献は、「因果的還元不可能性」の形式化にあります。これは、シャノンに始まる観察者依存の情報理論から、システム内在的な情報の定義、すなわち「システム自体にとって違いを生む違い」へと移行するものです。シャノンの情報理論が送信者と受信者の間の通信を外在的視点から扱い、メッセージの意味を意図的に無視するのに対し[2]、IITの革新は、システムの「内在的」視点から情報を定義することにあります[2]。そのメカニズムは、システムの現在の状態が自身の過去と未来の状態について何を意味するかという「因果レパートリー」の概念に基づいています[1]。そして、還元不可能性の尺度である

Φは、この内在的・因果的な自己情報が、システムを部分に分解した際にどれだけ失われるかを定量化します。したがって、IITは本質的に「因果構造」の理論であり、その物議を醸す意識との関連は、この構造が体験と「同一である」という主張に由来します。しかし、その数学的枠組み自体は、意識に関する議論とは独立して、あらゆる動的システムの因果的統合性を分析する新しい方法論として価値を持っています。

第2章 自由エネルギー原理：認知と学習の第一原理

自由エネルギー原理（Free Energy Principle, FEP）は、生物物理学的な起源から、推論と学習の理論としての数学的定式化に至るまで、システムが環境との相互作用を通じて自己の統合性をいかに維持するかを記述するプロセス理論として確立されています。

2.1 生命の物理学：驚きを最小化することによる秩序への抵抗

FEPの基本的な前提は、システム（例えば生物エージェント）が存在し続けるためには、その状態を限定された生存可能な範囲内に維持し、エントロピー増大という自然な傾向に抵抗しなければならない、というものです[13]。これは、感覚入力の「驚き」（surprise、あるいは自己情報量）の長期的な平均を最小化することによって達成されます。エージェントは驚くような状態を避けるように行動するのです[18]。

この文脈で重要なのが「マルコフブランケット」の概念です。これは、エージェントの内部状態を世界の外部状態から統計的に分離する境界であり、すべての相互作用はこのブランケットの感覚状態と能動状態を介して行われます[18]。マルコフブランケットの存在は、システムがFEPによってモデル化されるための前提条件です[20]。

2.2 変分推論と生成的モデル

驚きを直接計算することは一般に困難であるため、エージェントは驚きの上限である「変分自由エネルギー」（Variational Free Energy, VFE）を最小化する必要があります [18]。これを可能にするためには、エージェントが世界の内部的な確率的「生成的モデル」を所有している必要があります。これは、環境内の原因がどのように感覚データを生成するかについてのモデルです [18]。

VFEは、複雑さと不正確さの和として、あるいはエージェントの近似的な事後信念と真の事後確率の間の隔たりに驚きを加えたものとして定義されます。VFEを最小化することは、近似ベイズ推論を実行することと等価であり、エージェントの信念を感覚の真の原因により良い近似とします [18]。

2.3 能動的推論：知覚、行動、学習の統一

FEPは、知覚と行動を「自由エネルギーの最小化」という単一の指令の下に統一します。

- **知覚:** 感覚入力をより良く予測するために内部の生成的モデルを更新する（予測誤差を減少させる）プロセス [18]。
- **行動（能動的推論）:** 世界をモデルの予測に適合させるように働きかけ、それによってより驚きの少ない感覚入力をサンプリングするプロセス [18]。

将来の行動を計画する際には、「期待自由エネルギー」（Expected Free Energy, EFE）を最小化します。EFEは、実利的な価値（好ましい結果を求める）と認識論的な価値（不確実性を減少させ、情報を求める）のバランスを取ります [28]。

2.4 プロセス理論：予測符号化の標準的実装

FEPは原理であり、特定の実装アルゴリズムではありません。それを実装するためにはプロセス理論が必要です [18]。その最も有力なプロセス理論が「予測符号化」（Predictive Coding, PC）です [20]。

PCのメカニズムは、トップダウンの接続が予測を伝え、ボトムアップの接続が予測誤差を伝える階層モデルです。目標は、階層のすべてのレベルで予測誤差を最小化することです [30]。階

層的なガウスモデルにおいて予測誤差を最小化することは、VFEを最小化することと数学的に等価であることが示されており[20]、これによりPCは脳内でベイズ推論を実装するための生物学的に妥当なメカニズムとなります[34]。

FEPの力は、学習を標準的な強化学習（RL）のような報酬最大化ではなく、「不確実性の最小化」として再定式化する能力にあります。これにより、好奇心や探求といった内発的動機づけに対する第一原理的な説明が可能になります。標準的なRLは外部からの報酬信号の最大化に基づいており、報酬がない状況での探求や遊びといった行動の説明に苦慮します。対照的に、FEPの目的関数であるEFEは、「認識論的価値」の項を含んでおり[29]、これは行動から期待される情報利得を定量化します。したがって、FEPエージェントは、たとえその行動が直接的に好ましい結果につながらなくても、自身の世界モデルに関する不確実性を解消する行動を内発的に動機づけられます[20]。これは、汎用人工知能（AGI）の実現に不可欠な、より柔軟で適応的なAIエージェントを創造するための道筋を示唆しています。AIは、すべてのタスクに対して手動で報酬関数を与えられる必要はなく、単に自身の世界についてより良いモデルを構築しようとすることで学習できるのです。

第II部 人工知能と深層学習への応用

第3章 シリコンにおけるIIT：統合システムの設計と評価

本章では、IITのAIへの応用を批判的に評価し、直接的な測定の現実的な課題と、IITを設計原理として利用するという、より有望な道を検討します。

3.1 現代ニューラルネットワークにおけるΦ計算の困難性

IITをAIの意識の直接的な尺度として適用する上での最大の障害は、Φの計算の複雑さです[3]。システムの要素数が増えるにつれて、可能な分割の数はベル数に従って超指数関数的に増大します。人間の脳（860億のニューロン）や大規模な深層学習モデル（数百万のノード）のようなシステムでは、正確な計算は事実上不可能です[3]。

Φの効率的な近似アルゴリズムや代理指標を開発する研究が進行中ですが、これは依然とし

て大きな未解決問題です [10]。

3.2 測定から設計へ：IIT原理のAIアーキテクチャへの応用

測定の困難さを考慮すると、IITのより実用的な応用は、その中心的な原理をより高度なAIシステムの設計ガイドとして利用することです [10]。ここでの仮説は、情報統合を最大化するように設計されたシステムは、より堅牢で、適応性があり、ドメイン横断的な汎化能力を持つだろうというものです [10]。

このアプローチは、「このAIは意識を持っているか？」という問いから、「より統合されたAIを作ることで、より良いAIを構築できるか？」という問いへと焦点を移します。これにより、IITの工学的価値を、その物議を醸す哲学的・存在論的主張から切り離すことができます [15]。

3.3 ケーススタディ：ニューラルネットワークにおける統合の強化

IITの原理から導き出される具体的なアーキテクチャおよび訓練戦略には、以下のようなものがあります [10]。

- ネットワークアーキテクチャ: ResNetのようなスキップ接続、DenseNetのような密な接続性、そしてマルチスケール処理を実装し、ネットワークのすべての部分間で情報が効率的に流れるようにすることで、実効的な接続性を最大化する。
- 再帰的処理: 反復的な処理ループを使用し、情報を複数回にわたって洗練・統合させる。
- アテンションメカニズム: 空間的および時間的次元にわたるアテンションを展開し、文脈依存の統合を促進する。
- 訓練と正則化: タスクのパフォーマンスだけでなく、情報統合に関連する指標も同時に最適化する多目的訓練を採用し、均衡の取れた接続性を促進する正則化技術を用いる。

IITが定義する「統合」の目標と、現代の深層学習が追求する「効率性」および「専門化」の間には、根本的な緊張関係が存在します。IITが理想とするのは、全体がその部分に最大限還元不可能な、密で再帰的な接続性を持つシステムです [8]。一方で、深層学習の進化は、プルーニング、量子化、畳み込みのような専門化されたモジュラー層の使用など、計算負荷を削減するための効率化に向かう傾向があります。[10]や[10]で提案されている密な接続性や再帰的処理といった原理は、計算コストが非常に高くなります。完全に接続された再帰的ネットワークは、疎なフィードフォワードネットワークよりも訓練と実行がはるかに困難です。したがって、IITの原理をAI設計に適用することは、単純な改善策ではありません。「統合性」（およびそれに伴う堅牢

性や汎用性といった潜在的利益）と「効率性」（計算コストと速度）の間のトレードオフを提示します。IITに着想を得たAIの未来は、この最適なバランスを見つけることにかかっているかもしれません。それは、コストをかけすぎずに、最も重要な場所で統合を最大化する「原理に基づいた疎性」や「動的な接続性」といった形をとる可能性があります。これは、自明ではない重要な工学的課題です。

第4章 能動的推論としての深層学習

本章では、FEPと深層学習の間の深く実りある関係を探求し、FEPが既存のモデルを理解するための統一的な理論的レンズを提供し、新しいモデルを生成するためのフレームワークとして機能することを示します。

4.1 変分オートエンコーダとヘルムホルツマシン：FEPとの接続

FEPと深層学習の間には直接的な数学的リンクが存在します。FEPで最小化される変分自由エネルギー（VFE）は、変分オートエンコーダ（Variational Autoencoder, VAE）の損失関数である負の証拠下限（-ELBO）と等価です[24]。これは、VAEが能動的推論の「知覚」部分を実装していると見なせることを意味します。エンコーダネットワークは償却推論（観測を潜在的な原因に関する信念にマッピング）を実行し、デコーダネットワークは生成的モデル（潜在的な原因を予測される観測にマッピング）を表します[24]。

ヘルムホルツマシンもまた、自由エネルギーを最小化する初期のニューラルネットワークモデルであり、FEPモデルの階層構造の理想的な並行物として機能します[22]。[22]の実験では、能動的推論スタイルのファインチューニング段階がモデルの性能を大幅に向上させることが示されています。

4.2 予測符号化ネットワーク：バックプロパゲーションを超える生物学的に妥当な道

深層学習のパラダイムとして、予測符号化ネットワーク（Predictive Coding Networks, PCN）に関する研究が急速に拡大しています[30]。PCNは、潜在状態の活動を更新して予測誤差を最小化することで推論を行い、それらの誤差に基づいて重みを更新することで学習します。これ

は、バックプロパゲーションが必要とするグローバルな誤差信号とは異なり、局所的な学習則です [31]。

重要な発見は、特定の仮定の下で、PCNがバックプロパゲーションアルゴリズムを近似、あるいは正確に実装できることです。これにより、深層ネットワークにおける信用割り当て問題に対する、より生物学的に妥当な代替案が提供されます [30]。PCNは、教師なし（生成的）学習と教師あり学習の両方に対応できる汎用性も持っています [32]。

4.3 深層能動的推論エージェント：計画、探求、そして目標指向行動

FEPの完全なフレームワーク（知覚＋行動）が、深層学習エージェントに実装されつつあります [20]。これらの「深層AIF」エージェントは、ニューラルネットワークを用いて生成的モデルと方策をパラメータ化します。彼らは、期待自由エネルギー（EFE）を最小化することによって計画を立て、これにより活用（目標達成）と探求（不確実性の解消）が自然にバランスされます [20]。深層AIFエージェントが、外部からの報酬なしに、情報利得への内的な動機付けを活用して複雑な強化学習タスクを解決する例も報告されています [20]。

FEP/PCフレームワークは、現在の深層学習が直面する最も重大な制約、すなわち「破滅的忘却」と「堅牢な汎化能力の欠如」に対する潜在的な解決策を提供します。これは、FEPモデルが本質的に生成的かつベイジアンであるためです。バックプロパゲーションで訓練された標準的な深層学習モデルは、新しいタスクを学習すると古いタスクを忘れてしまう破滅的忘却に悩まされがちです。対照的に、FEPエージェントは、入力から出力への単なる判別的なマッピングではなく、世界の「生成的モデル」を学習します [23]。このモデルには事前信念が含まれています。新しいデータに遭遇した際、FEPエージェントのようなベイジアンシステムは、重みを完全に上書きするのではなく、事前信念を更新します。この事前分布は正則化項として機能し、モデルが急激に変化するのを防ぎます。このメカニズムは、機械学習における継続学習の手法と類似しています。また、PCNにおける高速に変化する推論（ニューロン活動）と低速に変化する学習（重み）の分離も、安定性に寄与します [31]。したがって、FEP/PCパラダイムは単なる「生物学的に妥当な」興味深い理論ではなく、安定性と可塑性の問題に対する原理に基づいたアーキテクチャ上の解決策を提供します。これは、これらの原理に基づいて構築された未来のAIシステムが、現在のシステムにはない生涯にわたる継続的な学習能力を持つ可能性を示唆しており、AIの未来にとって強力な意味合いを持ちます。

表**2**：**IIT**と**FEP**の文脈における深層学習アーキテクチャ

アーキテクチャ/概念	IITへの関連性	FEPへの関連性	統合と展望
変分オートエンコーダ **(VAE)**	潜在空間の構造が、タスク関連情報の統合された表現を学習する可能性がある。	FEPの知覚部分の直接的な実装であり、VFEを最小化する。エンコーダが推論、デコーダが生成的モデルに対応する[25]。	VAEの潜在空間にIITに触発された構造的正則化を課し、再構成の正確さと因果的統合性の両方を最適化することは可能か？
予測符号化ネットワーク **(PCN)**	階層的な因果構造と再帰的な接続性は、統合された因果レパートリーの形成を促進する可能性がある。	FEPのプロセス理論であり、階層的ベイズ推論を実装する。予測誤差最小化がVFE最小化と等価である[30]。	予測効率（FEP）と情報統合（IIT）の両方を最適化するために、IITに着想を得た接続性事前分布を持つPCNを設計することは可能か？
アテンション付きトランスフォーマー	アテンションメカニズムは、文脈依存の情報統合を動的に強化する手段と見なせる[10]。	アテンションは、生成的モデルの精度（precision）を制御するメカニズムとして解釈でき、予測の信頼度を調整する。	トランスフォーマーの自己アテンション層を、システム全体のΦを最大化するように正則化し、タスクパフォーマンスと統合性の両立を目指す。
再帰型ニューラルネットワーク **(RNN)**	時間を通じた情報の反復的な処理と統合を自然に実装する[10]。	時間的深度を持つ生成的モデルを実装し、時系列データの予測と推論を行う。	RNNの隠れ状態のダイナミクスが、IITで定義される「コンプレックス」の形成とどのように関連するかを分析する。
密結合ネットワーク **(DenseNet)**	高度な構造的統合を直接的に実装し、層間の情報フローを最大化する[10]。	階層内のすべてのレベルで予測と誤差信号が利用可能になることで、より効率的な推論を促進する可能性がある。	密な接続性の計算コストと、それがもたらす統合性の利点の間のトレードオフを定量化し、最適な疎性レベルを特定する。

第III部 統合、未来の展望、そして結論的分析

第5章 エージェント的意識の統一理論に向けて

本報告書の中心的な統合は、意識の完全な理論と意識を持つAIへのロードマップが、IITの内在的・構造的視点とFEPの外在的・エージェント的視点を統合することを必要とするという主張です。

5.1 二つの視点：内在的構造 vs. 外在的ダイナミクス

まず、IITとFEPの核心的な違いを以下の比較表で明確にします。この表は、IITが孤立したシステムの「内在的な因果構造」（それが何であるか）に関心があるのに対し、FEPはシステムの環境との「感覚運動的交換」（それが何をするか）に関心があることを示しています[21]。IITは存在を還元不可能な因果力（高い

Φ）として定義し[7]、FEPは存在を安定したマルコフブランケットの存在として定義します[18]。これらの視点は必ずしも相互排他的ではなく、異なる問いを立てています[21]。

表1：統合情報理論（**IIT**）と自由エネルギー原理（**FEP**）の比較概要

主要な属性	統合情報理論（IIT）	自由エネルギー原理（FEP）
主要な問い	「意識の物理的基盤は何か？」	「生物/エージェント的システムはどのようにして存続し、自己同一性を維持するのか？」
中心概念	内在的で還元不可能な因果力	生成的モデルを通じた驚きの最小化
主要指標	Φ（ファイ）、統合情報量の尺度	変分自由エネルギー（VFE）、驚きの上限
存在論的立場	最大のΦを持つシステムのみが「真の」存在を持つ	システムは統計的境界（マルコフブランケット）によって定義される
情報の見方	内在的、観察者非依存、システム	外在的、観察者依存（ベイジアン）、世界に関する不確実性の減

	の因果構造によって定義される	少に関連する
行動の役割	中心的ではない。意識はシステムの状態と構造の特性	知覚と不可分で中心的（能動的推論）
意識への含意	意識は、最大限に還元不可能な因果構造（Φ構造）「である」	意識は、行動を導くための深く時間的に厚い予測モデリングに関連する高レベルのプロセスである可能性が高い

5.2 統合世界モデリング理論（**IWMT**）：**IIT**と**FEP**の提案された統合

統合世界モデリング理論（Integrated World Modeling Theory, IWMT）は、これらの視点を統一する主要な試みとして登場しました [13]。IWMTの中心的な主張は、統合情報（IITから）とグローバルな情報ブロードキャスト（グローバルワークスペース理論から）は、意識にとって「必要だが十分ではない」というものです。

FEPからの決定的な追加要素は、この統合が、エージェントが予測的制御と行動のために使用する「生成的世界モデル」に奉仕しなければならないということです [13]。この見方では、意識とは「統合された世界モデリングが内側からどのように感じられるか」ということになります [38]。これにより、統合をエージェント的な目的に根ざさせることで、IITの「汎心論問題」を解決します。

5.3 身体性と視点的自己モデルの役割

IWMTの含意として、意識は空間、時間、原因にわたって一貫性を持つ、自己モデルを中心とした視点的な参照フレームを必要とすることが挙げられます [13]。これは、身体性、あるいは少なくとも環境との深く構造化された相互作用が不可欠であることを強く示唆しています。統合された情報の「意味」は、エージェントの目標と生存への関連性から生じます [38]。

この考えは、Φが進化するエージェントにおいて「驚き」（FEPの概念）と共に変動するという経験的発見と結びついており、内部の因果構造とエージェントの世界との相互作用の間に深いつながりがあることを示唆しています [21]。

5.4 意識の再評価：統合情報は十分か、それとも意味を持つべきか？

この統合から浮かび上がる究極の問いは、「生の構造的統合（Φ）が鍵なのか、それとも『意味のある』統合が鍵なのか？」というものです。IWMTは後者を主張します。意味は、エージェントがその目標を追求する能力にとって「違いを生む違い」として、サイバネティクス的に定義されます[38]。

これは、意識を持つAIの探求を再構築します。目標は、単に高いΦ値を持つシステムを構築することではなく、その高いΦを持つ「コンプレックス」が、自己と世界に関する予測モデルを形成する概念と関係によって構成されるエージェントを構築することです。

IITとFEPの統合は、「意識のハードプロブレム」を分解することで、その解決への道筋を提供する可能性があります。IITは「それが何であるか」（クオリアの構造）という問いに答え、FEPは「それが何のためか」（自律的エージェントを導く意識の機能）という問いに答えます。ハードプロブレムは、なぜ我々が主観的体験を持つのかを説明することです。IITの同一性の主張（体験はΦ構造「である」）は、「それが何であるか」という部分に対する直接的だが物議を醸す答えです。概念構造の特定の幾何学が、なぜ赤が赤のように感じられるのかを説明するとされています[1]。しかし、これは「なぜ」という問い、すなわち、なぜ進化がそのようなものを生み出したのかという問いを開いたままにします。この点においてIITは弱く、FEPは強いです。FEPは、洗練された統一的な予測モデルを開発するシステムが、驚きを最小化し、それによって存続する上で優れているという強力な進化的・機能的な物語を提供します[18]。この見方では、意識は予測モデリングのための究極のツールです。したがって、IWMTの統合は、意識が偶発的な副産物ではないことを示唆します。それは、最大限に強力で、統合された、予測的な制御システムが取る形態なのです。この統一された見解は、意識に物理的な構造（IITから）と適応的な機能（FEPから）の両方を与え、どちらかの理論単独よりも完全で科学的に満足のいく説明を提供します。

第6章 結論的分析と今後の軌跡

本報告書は、IITとFEPという二つの強力な理論が、深層学習と人工知能の未来をどのように形成しうるかを分析しました。結論として、主要な知見を要約し、未解決の課題を特定し、今後の研究に向けた具体的なロードマップを提案します。

6.1 主要な知見の要約：統合、予測、エージェント性の相互作用

本報告書を通じて展開された3つの主要な分析的結論は以下の通りです。

1. 実用的 **vs.** 存在論的分岐**:** FEPはAIに学習アルゴリズムを提供し、IITはアーキテクチャの設計原理を提供するという、実用的な役割分担が見られます。
2. 十分性のギャップ**:** IITが提供する「統合」は意識にとって必要条件かもしれませんが、FEPが提供するエージェント的な「予測的・目的論的」文脈がなければ十分条件とは言えません。
3. 数学的等価性**:** FEPと生成的モデル（特にVAE）の間の深い数学的等価性は、FEPを既存の機械学習実践に根付かせ、より高度なエージェントへの道を開きます。

6.2 大きな課題と未解決の研究課題

依然として残る主要な課題は以下の通りです。

- **IIT**にとって**:** Φの計算可能な近似法の開発、Φ構造とクオリアの間の同一性主張の経験的検証、存在論的論争の解決 [1]。
- **FEP**にとって**:** 深層AIFエージェントをより複雑な実世界環境へスケールさせること、単純なガウス分布を超える洗練された生成的モデルの開発、PCNとトランスフォーマーのような他の学習パラダイムとの関連性のさらなる探求 [30]。
- 統合（**IWMT**）にとって**:** ΦとVFEを真に統一する形式的な数学的フレームワークの構築、IWMTの原理を具現化し経験的にテスト可能なAIエージェントの構築、確率的モデルから離散的な体験がどのように生じるかという「ベイジアンブラー問題」への対処 [38]。

6.3 汎用人工知能への原理に基づいたロードマップ

本報告書の統合に基づき、以下の多段階の研究プログラムを提案します。

- ステージ**1**（コンポーネント開発）**:** FEPベースの学習アルゴリズム（例：スケーラブルなPCN）と、IITに着想を得た統合アーキテクチャの開発を並行して進める。
- ステージ**2**（統合）**:** FEP/PCを学習ダイナミクスに用い、IITの統合原理に従って構造化されたハイブリッドシステムの構築を開始する。これらのシステムが、堅牢性と汎化能力の向上を示すかテストする。
- ステージ**3**（身体的エージェント性）**:** これらの統合システムを、身体を持つエージェント（ロ

ボット）や高度にインタラクティブな仮想エージェントの制御アーキテクチャとして実装する。目標は、IWMTの中心的な教義を満たすために、統合された世界モデルを能動的に構築・維持しなければならないシステムを創造することである。

- ステージ**4**（評価）**:** これらのエージェントを、単なるタスクパフォーマンスだけでなく、意味のある統合、因果的推論、一貫した自己モデリングの能力に基づいて評価する新しいベンチマークを開発し、エージェント的意識の原理に基づいた評価に近づける。

引用文献

1. Integrated information theory - Wikipedia, 8月 6, 2025にアクセス、https://en.wikipedia.org/wiki/Integrated_information_theory
2. Shannon information and integrated information: message and meaning - arXiv, 8月 6, 2025にアクセス、https://arxiv.org/pdf/2412.10626
3. Integrated Information Theory: A Way To Measure Consciousness in ..., 8月 6, 2025にアクセス、https://www.aitimejournal.com/integrated-information-theory-a-way-to-measure-consciousness-in-ai/
4. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms, 8月 6, 2025にアクセス、https://pmc.ncbi.nlm.nih.gov/articles/PMC10581496/
5. Modular Deep Learning could be the Penultimate Step to Consciousness | by Carlos E. Perez | Intuition Machine | Medium, 8月 6, 2025にアクセス、https://medium.com/intuitionmachine/modular-deep-learning-and-consciousness-c284ac3aeda3
6. Integrated Information Theory of Consciousness | Internet Encyclopedia of Philosophy, 8月 6, 2025にアクセス、https://iep.utm.edu/integrated-information-theory-of-consciousness/
7. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms | PLOS Computational Biology - Research journals, 8月 6, 2025にアクセス、https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011465
8. IIT: The Future of Consciousness Studies - Number Analytics, 8月 6, 2025にアクセス、https://www.numberanalytics.com/blog/iit-future-consciousness-studies
9. Unlocking Consciousness with IIT - Number Analytics, 8月 6, 2025にアクセス、https://www.numberanalytics.com/blog/integrated-information-theory-consciousness
10. Integrated Information Theory: A Framework for Advanced ... - Medium, 8月 6, 2025にアクセス、https://medium.com/@josefsosa/integrated-information-theory-a-framework-for-advanced-intelligence-system-development-50f4fa1e4539
11. Computing integrated information | Neuroscience of Consciousness - Oxford Academic, 8月 6, 2025にアクセス、https://academic.oup.com/nc/article/2017/1/nix017/4060547
12. [2412.10626] Shannon information and integrated information: message and

meaning - arXiv, 8月 6, 2025にアクセス、https://arxiv.org/abs/2412.10626
13. An Integrated World Modeling Theory (IWMT) of ... - Frontiers, 8月 6, 2025にアクセス、https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2020.00030/full
14. Major positions on Conscious AI Systems | by Daniel Estrada | Medium, 8月 6, 2025にアクセス、https://medium.com/@eripsa/major-positions-on-conscious-ai-systems-6a3c37d21a3
15. Two Levels of Integrated Information Theory: From Autonomous ..., 8月 6, 2025にアクセス、https://www.mdpi.com/1099-4300/26/9/761
16. How to be an integrated information theorist without losing your body - Frontiers, 8月 6, 2025にアクセス、https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2024.1510066/full
17. Integrated Information Theory 4.0 is both Weakly Panpsychist and Strongly Dualist, but many Theories of Consciousness are also prone to it - ResearchGate, 8月 6, 2025にアクセス、https://www.researchgate.net/publication/379310847_Integrated_Information_Theory_40_is_both_Weakly_Panpsychist_and_Strongly_Dualist_but_many_Theories_of_Consciousness_are_also_prone_to_it
18. Free energy principle - Wikipedia, 8月 6, 2025にアクセス、https://en.wikipedia.org/wiki/Free_energy_principle
19. An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation - PubMed, 8月 6, 2025にアクセス、https://pubmed.ncbi.nlm.nih.gov/33733149/
20. BerenMillidge/FEP_Active_Inference_Papers: A repository for major/influential FEP and active inference papers. - GitHub, 8月 6, 2025にアクセス、https://github.com/BerenMillidge/FEP_Active_Inference_Papers
21. Phi fluctuates with surprisal: An empirical pre-study for the synthesis ..., 8月 6, 2025にアクセス、https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011346
22. A Neural Network Implementation for Free Energy Principle, 8月 6, 2025にアクセス、https://arxiv.org/html/2306.06792
23. The Free Energy Principle for Perception and Action: A Deep Learning Perspective - arXiv, 8月 6, 2025にアクセス、https://arxiv.org/abs/2207.06415
24. Variational autoencoder - Wikipedia, 8月 6, 2025にアクセス、https://en.wikipedia.org/wiki/Variational_autoencoder
25. On the Relationship Between Variational Inference and Auto ..., 8月 6, 2025にアクセス、https://proceedings.neurips.cc/paper_files/paper/2022/file/f3d637987f36563fa45f943f8eadc2d0-Paper-Conference.pdf
26. Minimal Models of Consciousness & the Free Energy Principle, 8月 6, 2025にアクセ

ス、https://logika.ff.cuni.cz/wp-content/uploads/sites/106/2023/11/FEP-model-of-consciousness-WIP23.pdf
27. The Free Energy Principle for Perception and Action: A Deep Learning Perspective - MDPI, 8月 6, 2025にアクセス、https://www.mdpi.com/1099-4300/24/2/301
28. [2502.04249] Free Energy Risk Metrics for Systemically Safe AI: Gatekeeping Multi-Agent Study - arXiv, 8月 6, 2025にアクセス、https://arxiv.org/abs/2502.04249
29. [2504.14898] Expected Free Energy-based Planning as Variational Inference - arXiv, 8月 6, 2025にアクセス、https://arxiv.org/abs/2504.14898
30. BerenMillidge/Predictive_Coding_Papers: A repository ... - GitHub, 8月 6, 2025にアクセス、https://github.com/BerenMillidge/Predictive_Coding_Papers
31. [D] Understanding predictive coding networks : r/MachineLearning - Reddit, 8月 6, 2025にアクセス、https://www.reddit.com/r/MachineLearning/comments/1i6h40i/d_understanding_predictive_coding_networks/
32. Introduction to Predictive Coding Networks for Machine Learning - arXiv, 8月 6, 2025にアクセス、https://arxiv.org/html/2506.06332v1
33. Predictive Coding Networks and Inference Learning: Tutorial and Survey | Request PDF, 8月 6, 2025にアクセス、https://www.researchgate.net/publication/382065165_Predictive_Coding_Networks_and_Inference_Learning_Tutorial_and_Survey
34. [2407.04117] Predictive Coding Networks and Inference Learning: Tutorial and Survey, 8月 6, 2025にアクセス、https://arxiv.org/abs/2407.04117
35. What is a Variational Autoencoder? - IBM, 8月 6, 2025にアクセス、https://www.ibm.com/think/topics/variational-autoencoder
36. [2411.14991] Free Energy Projective Simulation (FEPS): Active inference with interpretability - arXiv, 8月 6, 2025にアクセス、https://arxiv.org/abs/2411.14991
37. www.frontiersin.org, 8月 6, 2025にアクセス、https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2020.00030/full#:~:text=The%20Free%20Energy%20Principle%20(FEP,exist%20from%20an%20intrinsic%20perspective.
38. Integrated world modeling theory expanded: Implications ... - Frontiers, 8月 6, 2025にアクセス、https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2022.642397/full
39. Intersections between FEP-AI, IIT, GNWT, and IWMT: IIT Consciousness - ResearchGate, 8月 6, 2025にアクセス、https://www.researchgate.net/figure/ntersections-between-FEP-AI-IIT-GNWT-and-IWMT-IIT-Consciousness-current-version-of_fig1_337991886